DOCUMENT RESUME

ED 210 273 TM 810 670

AUTHOR Donlon, Thomas F.
TITLE Judged Similarity of Aptitude and Achievement Tests in Mathematics.
PUB DATE 81
NOTE 10p.; Paper presented at the New England Educational Research Organization (Lenox, MA, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Aptitude Tests; Comparative Analysis; *Item Banks; *Quantitative Tests; Secondary School Mathematics; *Specifications; *Test Format; Test Items
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

This study attempts to establish the ability of a panel of five judges with varied mathematics background to distinguish between two types of mathematical tests by separating their component items when they are presented in a mixed pool of aptitude and achievement tests. Typically, the two tests show high correlation. The judges showed about 70% success in their judgments, which points to a clear distinction between the tests. In examining the items which were consistently misclassified, the general conclusions were that certain achievement items exhibit aptitude-item characteristics, and vice-versa. The use of more formal specifications might be useful in controlling the correlation between the tests. However, as mathematics curricula introduce formerly advanced topics at earlier times, such specifications will probably become increasingly difficult to establish. (Author/AL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED210273

JUDGED SIMILARITY OF APTITUDE AND ACHIEVEMENT TESTS IN MATHEMATICS

Thomas F. Donlon
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T.F. Donlon
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810670

Paper presented at the New England Educational Research Organization, Lenox, Massachusetts, 1981.

Judged Similarity of Aptitude and Achievement Tests in Mathematics

Thomas F. Donlon
Educational Testing Service

There has been limited work to date on the extent to which test classifications of various kinds are rationally supported by consistent item classifications. Cook (1963) investigated the extent to which classroom teachers could correctly identify items as deriving from tests of intelligence, aptitude, or achievement. His hypotheses were that: (1) the proportion of items assigned by the teachers to the three categories would be random and hence approximately equal; and that (2) the proportion of teachers correctly identifying each type of item would be equal to chance. These hypotheses were disproved, and he concluded that the teachers in his group did have the ability through their training and experience to correctly differentiate between the three types of items. The achievement type items were the easiest for them to differentiate, while the aptitude items were the hardest.

Stafford (1974) attempted to replicate Cook's findings using psychology students in an assessment course as subjects instead of classroom teachers. The questions asked in this study were (1) will the different type of student make any difference?, (2) are students more "test sophisticated" in 1974 than in 1963?, and (3) are certain kinds of items more difficult to identify than others?

The same test items used by Cook were used by Stafford. These consisted of 20 intelligence items taken from the Otis Gamma, 20 aptitude items taken from the DAT and 20 achievement items taken from the Essential High School Content Battery. The items were randomly distributed and presented as one 60-item test.

The students were instructed to read the item and then mark their answer sheet according to which of the three types of tests they believed the item belonged. The subjects were 44 undergraduate juniors and seniors, primarily psychology majors (both male and female) taking a course in Psychological Assessment. The results indicated that the majority of students were able to correctly identify 34 out of 60 items at the 5% level of significance or better. They were best able to identify achievement test items with 18 correct, while they correctly identified aptitude and intelligence test items with 8 items correct each.

Commenting on the study, Stafford observed.

> "...The three kinds of tests contain items of highly similar nature. Perhaps one should ask, when given a clear definition of an aptitude test vs. an intelligence test or an achievement, why could not all students correctly identify all items? This seemingly overlapping of items types is, of course, the cause of such high correlations between these three types of tests and makes identification so difficult.

It would seem important to reduce this "test tautology" by designing testing instruments in such a unique manner that, given a reasonable verbal explanation of the three types, Psychology students in an assessment class and classroom teachers could correctly identify nearly all the items."

The inclusion of two mathematics instruments is a frequent characteristic in test batteries used in educational decision-making. Often, one of these is considered to assess mathematical aptitude, the other mathematical achievement. The specifications for the two tests are typically distinct, stressing the degree of relation to a curriculum or curriculum, the correlations between the measures they yield, however, can be substantial, on the order of .75 or more. Nonetheless, the concept of an aptitude measure, as somewhat curriculum free, is a persistently attractive one, and the use of separate tests continues.

The Present Study

While high correlations cannot necessarily be taken to mean that two tests are equivalent, they raise the question of the merits of distinctions between them. This study attempted to establish the ability of judges to distinguish between two types of mathematical tests by separating their component items when these were presented in a mixed pool of aptitude and achievement material.

The SAT-M is a test of mathematical achievement which is intended to measure the potential for further mathematical work. It measures very basic achievement, reflecting learning from the elementary grades. It is the product of the Mathematics Department at ETS. For years, it has been offered in conjunction with two levels of achievement test, of which the lower level is the subject of this study.

The two tests typically show a high correlation -- in scaling the achievement test over a number of administrations, the values observed for the correlation between various forms of the SAT and various forms of the achievement test, range in the low .80s.

How different are these two tests? How great is the overlap between them? While high correlation is not necessarily an indication of similar content, do the specifications indeed maintain a separateness between these two tests?

This study attempted to answer these questions by asking judges to identify the source of the items in a pool containing both SAT-M items and mathematics achievement items. If the two tests were perfectly distinct, the judges should be able to correctly distinguish among the items; to the extent that the tests overlap, the judges would classify actual SAT-M items as achievement items and vice-versa.

There were five judges in all. Two were members of the Mathematics Department; two were professional staff members at ETS who majored in Mathematics as collegiate undergraduates. The fifth judge was a college mathematics instructor. All judges were told to base their decisions on the descriptions of the tests which were provided to the candidates via the booklets published by the Board. The two Mathematics Department people, of course, could not help but have a more complex and developed conception of the two tests and the distinctions between them.

The pool consisted of 98 items. There were 48 aptitude items, which constituted one form of SAT-M with the exception of certain item material which had a particular format (Data Sufficiency) and which items did not appear in the achievement test. The 50 achievement items constituted one form of this test. Only one judge, one of the Mathematics Department members, was aware of the 50-48 split or that the items of one type were all from one single test of that type. The other judges were simply told that it was a random collection. Each of the judges from the Mathematics Department recognized certain items and excluded them from the sorting. One judge excluded 2 items on this basis, the other excluded 9 items. The pool was arranged as follows: The aptitude items were placed in order from 1 to 48; the achievement items from 1 to 50, and a coin was tossed to determine from which of these piles item 1 in the pool would come. This process was repeated until the piles were completely merged. This method has some advantage in that order within these two tests is roughly coordinated with difficulty, so that the more difficult aptitude items were presented with the more difficult achievement items.

The results of the sorting by each of the five judges are shown in the five contingency tables grouped together as Table 1 and presented below:

Table 1

Judge A

		Actual: Apt	Actual: Ach	
Judge: Ach		3	26	29
Judge: Apt		45	24	69
		48	50	98

Judge B

		Actual: Apt	Actual: Ach	
Judge: Ach		17	36	53
Judge: Apt		31	14	45
		48	50	98

Judge C

		Actual: Apt	Actual: Ach	
Judge: Ach		22	43	65
Judge: Apt		24	7	31
		46	50	96

(2 excluded as known)

Judge D

		Actual: Apt	Actual: Ach	
Judge: Ach		22	40	62
Judge: Apt		26	10	36
		48	50	98

Judge E

		Actual: Apt	Actual: Ach	
Judge: Ach		12	35	47
Judge: Apt		31	11	42
		43	46	89

(9 excluded as known)

On the average, the judges made about 30% misclassifications. The best judge had 26% errors, the worst, 33%. The judges varied somewhat in the criteria used to pick the classes. Thus, Judge A was conservative in defining the Achievement category. Judges C and D, on the other hand, appear to have been more conservative in defining their Aptitude categories. The other judges show more balance, but Judge E was the Mathematics Department member who was aware of the balance in the source of the pool.

The evidence of these tables, individually and collectively, supports the notion that the two tests are reasonably distinct. The judges had, in general, a minimum of preparation for this task, and the fact that they could be so successful when guided only by the simple test descriptions given to the candidates, suggests that an even greater accuracy could be achieved by a more thorough preparation.

On the other hand, the judges were all fairly capable as mathematicians. Certainly they could solve the items and they knew, in general, when the mathematics involved in a given question was likely to be presented in a high school curriculum. To what extent should the subject matter for these two tests be so distinct that judges would make only rare misplacements? Wouldn't it be desirable to have no overlap between them?

To facilitate a consideration of these questions, the study looked to see which items, if any, the judges had consistently misplaced.

The items were classed in accordance with the number of judges who categorized the item correctly as to test source. Thus, the items could vary from 0, or no judges guessing correctly, to 5, with all judges correctly identifying the test which was the source. There were 11 cases of items without ratings due to judges' recognition of the items. In these cases the missing rating was arbitrarily considered to agree with the majority of the ratings given by the other judges.

The results are presented below in Table 2.

Table 2

The Items Classed by Number of Judges Correctly Identifying Test Source

No. of Judges	Aptitude Test	Achievement Test
5	13	19
4	14	13
3	7	9
2	8	1
1	6	5
0	0	3
	48	50

The data in Table 2 indicate that there was a substantial agreement among the judges, with only 25 of the 98 items receiving an ambiguous split in which 2 of the judges vote one way and 3 vote the other way. They also indicate that 6 of the aptitude items and 8 of the achievement items were consistently attributed to the wrong test by 4 or more of the judges. Among those aptitude items which the judges considered to be achievement items, 5 involved formal algebra; the other, presented here as Item A in Figure 1, involved the use of the Pythagorean theorem. Of the Achievement Test items which were classed "aptitude" by the judges, 5 were apparently so easy as to fall into the content domain of the SAT-M.

Figure 1 presents the three achievement items which were reasonably difficult but which were judged as "aptitude" by all five judges. Item B, which is straightforward logic, was apparently perceived by the judges as too nonmathematical to be considered achievement. Items C & D, on the other hand, draw upon relationships derived from courses, but apparently present the material with a sufficiently unique "aptitude" twist to win separate status.

Summary and Conclusions

When asked to identify the source test in a mixed pool of SAT-M items and Intermediate Mathematics items, a panel of 5 judges with varied mathematics background showed about 70% success in their judgments. There were tendencies for individual judges to establish more stringent criteria for one or the other of the two categories. Thus, one judge singled out only 29 items as "achievement" but was "right" in 26 of these cases.

This evidence points to a clear distinction between the tests, even greater, perhaps, than would be anticipated on the basis of the correlations between them. In examining the items which were consistently misclassified, the general conclusions were that certain aptitude items called for a more formal exercise in algebra, or for specialized knowledge of inequalities or the Pythagorean theorem, while some of the achievement items measured pre-secondary school achievements or were couched in a problem-format more typical of aptitude items.

The concept of a problem-format typical of the Aptitude Test is an important one for the Mathematics Department. The use of non-straightforward problems in achievement tests is something that has been consciously elected, not unconsciously drifted into, so that, when the judges "reject" certain achievement items as "aptitude" items, they are merely weighting the format or process dimension of the item more heavily than they are weighting the content dimension. The Mathematics Department readily recognizes the two dimensions. It may be, however, that a more formal specification of the proportion of achievement items with marked aptitude-item characteristics would be useful in controlling the correlation between the tests. At the same time, a more formal specification of the number of SAT-M items involving secondary-school mathematics might usefully support the content and construct validity of this instrument. As mathematics curricula increasingly introduce formerly advanced topics at earlier times, however, it probably becomes increasingly difficult to establish such specifications.

/gv

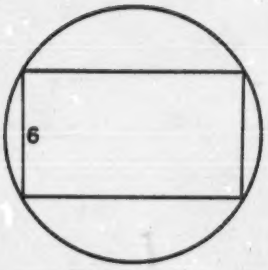
REFERENCES

Cook, D. L., Teachers' Ability to Identify Intelligence, Achievement, and Aptitude Test Items. Paper presented APA, Philadelphia, Pa., Sept. 2, 1963.

Stafford, R. E. The Ability to Differentiate Between Intelligence, Achievement and Aptitude Items. Paper presented at NCME, Chicago, Ill., April 18, 1974.

APPENDIX

Item A.



In the above figure a circle passes through the vertices of a rectangle. If the width of the rectangle is 6 and its area is 48, what is the area of the circle?

(A) $\frac{25\pi}{4}$ (B) 16π (C) 25π (D) 64π (E) 100π

Item B.

Given: If P is true, then Q is true. Which of the following is a valid conclusion?

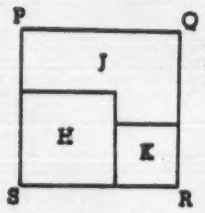
(A) If Q is true, then P is true.
(B) If P is true, then Q is not true.
(C) If Q is true, then P is not true.
(D) If P is not true, then Q is not true.
(E) If Q is not true, then P is not true.

Item C.

If the sides of a rectangle are integers and if the sum of three sides is 8, then what is the difference between the greatest and least possible areas?

(A) 1 (B) 2 (C) 3 (D) 4 (E) 5

Item D.



In the above figure PQRS is a square. If H and K are squares of area x^2 and y^2, respectively, what is the area of J?

(A) $x^2 + y^2$ (B) $2xy + y^2$ (C) $(x + y)^2$
(D) $2xy$ (E) $2\sqrt{xy}$